CLAIMS

1. A method for representing the latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

deriving at least one n-tuple term from the plurality of terms;

forming a two-dimensional matrix,

each matrix column c corresponding to a document,

each matrix row r corresponding to a term occurring in at least one document corresponding to a matrix column,

each matrix element (r, c) related to the number of occurrences of the term corresponding to the row r in the document corresponding to column c,

at least one matrix element related to the number of occurrences of one at least one n-tuple term occurring in the at least one document, and

performing singular value decomposition and dimensionality reduction on the matrix to form a latent semantic indexed vector space.

2. The invention as recited in Claim 1 further comprising:
identifying an occurrence threshold;

wherein n-tuples that appear less times in the document collection than the occurrence threshold are not included as elements of the matrix.

3. The invention as recited in Claim ~~1~~2 wherein the occurrence threshold is two.

4. The invention as recited in Claim 1 wherein deriving at least one n-tuple term further comprises:

creating the at least one n-tuple term from n consecutive verbatim terms.

5. A method for determining conceptual similarity between a subject document and at least one of a plurality of reference documents, each document containing a plurality of terms, the method comprising:

deriving at least one n-tuple term from the plurality of terms,

forming a plurality of two-dimensional matrices wherein, for each matrix:

each matrix column c corresponds to a document, one column corresponding to the subject document;

each matrix row r corresponds to a term occurring in at least one document corresponding to a matrix column,

each matrix element (r, c) represents the number of occurrences of the term corresponding to r in the document corresponding to c;

performing singular value decomposition and dimensionality reduction on a plurality of formed matrices, to form a plurality of latent semantic indexed vector spaces,

the latent semantic indexed vector spaces including at least one space formed from a matrix including at least one element corresponding to the number of occurrences of at least one n-tuple term in at least one document,

determining at least one composite similarity measure between the subject document and at least one reference document as a function of a weighted similarity measure of the subject document to the reference document in each of a plurality of indexed vector spaces.

6. The method as recited in Claim 5 wherein the similarity measures from vector spaces comprising greater numbers of n-tuples are weighted greater than similarity measures from vector spaces comprising lesser number of n-tuples.

7. A method for representing a query document, the query document containing verbatim terms, the query document intended for querying a collection of reference documents via a latent semantic indexed representation of the reference collection; the method comprising:

identifying verbatim entities;

stemming identified entities;

generalizing stemmed entities; and

supplementing verbatim entities with corresponding generalized entities.

8. A method for representing a query document, the query document containing verbatim terms, the query document intended for querying a collection of reference documents via a latent semantic indexed representation of the reference collection; the method comprising:

identifying verbatim entities;

stemming identified entities;

generalizing stemmed entities; and

replacing verbatim entities with corresponding generalized entities.

9. The method as recited in Claim 8 wherein verbatim entities comprise ordered terms between stop words.

10. The method as recited in Claim 8 wherein generalizing entities further comprises alphabetically ordering stemmed entities as an aid to generalization.

11. The method as recited in Claim 8 wherein generalizing entities further comprises ordering stemmed entities as a function of the frequency of occurrence of verbatim entities.

12. The method as recited in Claim 8 wherein generalized entities are identified with human feedback.

13. The method as recited in Claim 8 wherein generalized entities are identified by automated process.

14. A method for characterizing the results of a query into a latent-semantic-indexed document space, the query comprising at least one term, the results comprising a set of document identities; the method comprising:

ranking results as a function of at least the frequency of occurrence of at least one term.

15. The method as recited in Claim 14 wherein at least one term used in ranking is a query term.

16. The method as recited in Claim 15 wherein the at least one query term used in ranking is a generalized entity.

17. The method as recited in Claim 14 wherein the at least one term used in ranking is a generalized entity.

18. A method for determining conceptual similarity between a query document and at least one of a plurality of reference documents, each document comprising a plurality of verbatim terms, the reference documents indexed into a latent semantic index space, the method comprising:

identifying verbatim entities;

stemming identified entities;

generalizing stemmed entities;

replacing at least one verbatim entity with the corresponding generalized entity to form a generalized query;

identifying a set of reference documents based on closeness, within the latent semantic indexed space, between the generalized query and each reference document; and

ranking a subset of closest identified documents as a function of at least the frequency of occurrence of at least one term.

19. The method as recited in Claim 18 wherein at least one term used in ranking is a query term.

20. The method as recited in Claim 19 wherein the at least one query term used in ranking is a generalized entity.

21. The method as recited in Claim 18 wherein the at least one term used in ranking is a generalized entity.

22. A method for representing the latent semantic content of a plurality of documents, each document containing a plurality of verbatim terms, the method comprising:

deriving at least one expansion phrase from the verbatim terms,

each expansion phrase comprising terms;

replacing at least one occurrence of a verbatim term having an expansion phrase with the expansion phrase corresponding to that verbatim term;

forming a two-dimensional matrix,

each matrix column c corresponding to a document;

each matrix row r corresponding to a term;

each matrix element (r, c) representing the number of occurrences of the term corresponding to r in the document corresponding to c;

at least one matrix element corresponding to the number of occurrences of one at least one term occurring in the at least one expansion phrase, and

performing singular value decomposition and dimensionality reduction on the matrix to form a latent semantic indexed vector space.

23. A method for representing the latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

identifying at least one idiom among the documents,

each idiom containing at least one idiom term;

forming a two-dimensional matrix,

each matrix column corresponding to a document;

each matrix row corresponding to a term occurring in at least one document represented by a row;

each matrix element representing the number of occurrences of the term corresponding to the element's row in the document corresponding to element's column;

at least one occurrence of at least one idiom term being excluded from the number of occurrences corresponding to that term in the matrix,

performing singular value decomposition and dimensionality reduction on the matrix.

24. A method for representing the latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

identifying at least one idiom among the documents,

each idiom containing at least one idiom term;

replacing at least one identified idiom with a corresponding idiom elaboration, each elaboration comprising at least one elaboration term,

forming a two-dimensional matrix,

each matrix column corresponding to a document;

each matrix row corresponding to a term;

each matrix element representing the number of occurrences of the term corresponding to the element's row in the document corresponding to element's column,

at least one matrix element corresponding to the number of occurrences of an elaboration term in a document corresponding to a matrix column;

performing singular value decomposition and dimensionality reduction on the matrix.